1.

feature 1	output
$x_1^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$y^{(2)}$
⋮	
$x_1^{(m)}$	$y^{(m)}$

다음과 같이 데이터가 주어졌다고 생각해보자. feature과 output이 선형적으로 관계한다고 생각하면 다음과 같이 표현할수있다.

$$y \simeq \theta_0 + \theta_1 x_1$$

우리의 목표는 위의 값의 좌우가 최대한 일치하도록 θ_0, θ_1을 결정해야한다.

그러기 위해서는 다음의 cost-function의 값을 최소화시켜줘야한다.

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = J(\theta)$$

대부분의 값이 다르면 다를수록 $\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$이 커지기 때문이다.

이 $J(\theta)$값이 작아지도록 θ을 설정하기 위해 우리는 gradient descent를 사용한다.

gradient descent는 $J(\theta)$에 대하여 $\frac{\partial J(\theta)}{\partial \theta_i}$를 구하고, 만약 0보다 이 값이 작을경우 θ_i을 늘렸을때 값이 작아지므로 θ_i를 + 방향으로 움직이고 또한 0보다 클경우에는 θ_i을 늘렸을때 값이 커지므로 θ_i을 - 방향으로 움직인다. 그러므로 θ_i를 다음과 같이 갱신해야함을 알수 있다.

$$\theta_i \leftarrow \theta_i - \alpha\frac{\partial J(\theta)}{\partial \theta_i}$$

원래의 $J(\theta)$와 갱신된 θ들을 넣은 $J(\theta)$가 거의 다르지 않을경우 갱신을 종료한다.

2. multiple-variable의 경우 $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n$ 이된다.

∴ Cost function은

$$J(\theta) = J(\theta_0 \cdots \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} \cdots \theta_n x_n^{(i)} - y^{(i)})^2$$

이된다.

위와 같이 정의할수 있는 이유는 $x_0, x_1 \cdots x_n$이 y에 대해서 선형적으로 관여한다고 생각했기 때문이다. 따라서

$$y \simeq \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n = h_\theta(x)$$

가 되고 좌변과 우변차이가 가장작은 $\theta_0 \cdots \theta_n$을 구하면 되는 것이다.

그러기 위해서는 $|\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n - y|$ 를 작게 만들어야 한다.

하지만 절댓값의 미운은 조금복잡할수 있으므로 $(\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n - y)^2$을 가장 최소가 되도록 설정해주면된다. ∴ 위의 $J(\theta)$는 cost function에 적절한 함수임을 알수 있다.

3. $h_\theta(x)$와 y가 비슷해지기 위해서

cost function인 $\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2 = J(\theta)$가 최소가 되어야한다.

$J(\theta)$값이 작아지도록 θ을 설정하기 위해 우리는 gradient descent를 사용한다.

gradient descent란 $J(\theta)$에 대하여 $\frac{\partial J(\theta)}{\partial \theta_i}$를 구하고, 만약 0보다 이 값이 작을 경우 θ_i를 늘렸을 때 값이 작아지므로 θ_i를 +방향으로 움직여야한다. 또한 0보다 클 경우에는 θ_i를 늘렸을 때 값이 작아지므로 θ_i를 + 방향으로 움직여야 함을 알 수있다. 그러므로 θ를 다음과 같이 갱신해야 적절하다고 생각할 수 있다.

$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial J(\theta)}{\partial \theta_i}$$

← α값이 너무 작다면 수렴이 국소에서 비교적 작은 θ값을 갱신할 것이다.

원래의 $J(\theta)$와 갱신된 θ을 넣은 $J(\theta)$가 거의 다르지 않을 경우 갱신을 종료한다.